HCDR Project - Phase 3

Team 11: Neelan Scheumann, Sha Liu, Ketan Pimparkar

4 P’s Report

- Past
  - In Phase 0, we created an outline of the project and the team plan to tackle the problem.
  - In Phase 1, we did EDA and feature engineering on train and 4 additional datasets. We also evaluated baseline model using Logistic Regression,
  - In Phase 2, we used 96 features and trained ML models with Logistic regression,Random Forest with and without RandomSearch CV, XGBoost and LightGBM algorithms. Our highest Kaggle scores are 0.75106 on private board and 0.74922 on public board.
- Present
  - For Phase 3, we performed additional feature engineering on available datasets and included deep learning model.
  - We used 180 features from available datasets to train these models.
  - Our highest Kaggle scores are 0.77474 on private board and 0.78264 on public board.
- Proposed
  - If we could spend more time working on this dataset, we could improve the deep learning model to achieve better results.
  - Additional feature engineering and feature selection would also help us improve the performance of this model.
- Problems
  - RandomizedSearchCV on Random Forest, XGBoost, and LightGBM model need longer time for us to train.

Feature Engineering:
(180 total features)

1. We did feature engineering for the application train datasets and the 6 additional datasets by creating new features for each datasets
2. After we did feature engineering for the application train datasets and the 6 additional datasets, we came up 180 features for the merged training dataset. We selected top 100 highly correlated numerical features and then 16 categorical features.
3. In total we have 116 features before one hot coding for the categorical features.
4. After one hot coding for the categorical features, we have 318 features in total. Feature selection resulted in 180 features.
5. See our features in next slides

Our Initially Selected 116 Features

```
num_attribs = list(X_train.select_dtypes(exclude='object').columns)
print("Numerical Features are:")
print(num_attribs)
print("-------------------------")
print('Number of numberical features: ', len(num_attribs))
```

```
Numerical Features are:
['Total_Remaining_repays', 'PREV_APP_TYPE_XNA', 'Max_Initial_term', 'TOTAL_DEBT_OVERDUE', 'DEBT_TO_INCOME', 'PREV_APP_PORTFOLIO_Cash', 'TIME_MORN', 'FLAG_DOCS_SUBMITTE
D', 'WEEKDAY_START', 'PREV_APP_STATUS_Canceled', 'CNT_CHILDREN', 'WEEKDAY_MID', 'CLOSED_IN_LAST_YEAR', 'SK_ID_PREV', 'APPLIED_EXTRA_0', 'AMT_REQ_CREDIT_BUREAU_YEAR',
'Avg_installment_days_difference', 'PERCENT_CREDIT_CARD', 'CC_Average_Monthly_Payments', 'PREV_APP_TYPE_Cashloans', 'FLAG_WORK_PHONE', 'DEF_60_CNT_SOCIAL_CIRCLE', 'DEF
_30_CNT_SOCIAL_CIRCLE', 'PREV_APP_PORTFOLIO_XNA', 'TIME_EARLY_MORN', 'LIVE_CITY_NOT_WORK_CITY', 'EXCESS_LOAN', 'AMT_CREDIT - AMT_GOODS_PRICE', 'OWN_CAR_AGE', 'PREV_APP
_PORTFOLIO_Cards', 'DAYS_REGISTRATION', 'FLAG_DOCUMENT_3', 'REG_CITY_NOT_LIVE_CITY', 'PREV_APP_TYPE_Revolvingloans', 'FLAG_EMP_PHONE', 'DAYS_DECISION', 'EXT_SOURCE_ST
D', 'REG_CITY_NOT_WORK_CITY', 'DAYS_ID_PUBLISH', 'DAYS_LAST_PHONE_CHANGE', 'PERCENT_LATE', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', 'PREV_APP_STATUS_Refu
sed', 'CREDIT_TO_APP_RATIO', 'TOTAL_NUMBER_OF_ACTIVE_LOANS', 'DAYS_EMPLOYED / DAYS_BIRTH', 'DAYS_EMPLOYED', 'CC_Average_Monthly_Balance', 'OPENED_IN_LAST_YEAR', 'EXT_S
OURCE_MEAN', 'app EXT_SOURCE_2 * EXT_SOURCE_3', 'app EXT_SOURCE prod', 'app EXT_SOURCE_1 * EXT_SOURCE_3', 'EXT_SOURCE_3', 'app EXT_SOURCE_1 * EXT_SOURCE_2', 'EXT_SOURC
E_2', 'EXT_SOURCE_1', 'DAYS_BIRTH', 'CC_Credit_Cards_Count', 'NAME_EDUCATION_TYPE_higher_education', 'FLOORSMAX_AVG', 'FLOORSMAX_MEDI', 'FLOORSMAX_MODE', 'Count', 'AMT
_GOODS_PRICE', 'REGION_POPULATION_RELATIVE', 'AMT_ANNUITY_y', 'ELEVATORS_AVG', 'ELEVATORS_MEDI', 'FLOORSMIN_AVG', 'FLOORSMIN_MEDI', 'LIVINGAREA_AVG', 'LIVINGAREA_MED
I', 'FLOORSMIN_MODE', 'TOTALAREA_MODE', 'ELEVATORS_MODE', 'AMT_CREDIT/AMT_ANNUITY', 'PREV_APP_STATUS_Approved', 'LIVINGAREA_MODE', 'AMT_CREDIT_x', 'APARTMENTS_AVG', 'A
PARTMENTS_MEDI', 'FLAG_DOCUMENT_6', 'APARTMENTS_MODE', 'Avg_installment_amount_difference', 'LIVINGAPARTMENTS_AVG', 'LIVINGAPARTMENTS_MEDI', 'PREV_APP_PORTFOLIO_POS',
'INS_AMT_PAYMENT_SUM', 'HOUR_APPR_PROCESS_START', 'FLAG_PHONE', 'LIVINGAPARTMENTS_MODE', 'BASEMENTAREA_AVG', 'YEARS_BUILD_MEDI', 'YEARS_BUILD_AVG', 'BASEMENTAREA_MED
I', 'YEARS_BUILD_MODE', 'AMT_APPLICATION', 'PREV_APP_TYPE_Consumerloans']
-------------------------
Number of numberical features:  100
```

```
print("Categorical Features are:")
cat_attribs = list(X_train.select_dtypes(include='object').columns)
print(cat_attribs)
print("-------------------------")
print('Number of Categorical features: ', len(cat_attribs))
```

```
Categorical Features are:
['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUS
ING_TYPE', 'OCCUPATION_TYPE', 'WEEKDAY_APPR_PROCESS_START', 'ORGANIZATION_TYPE', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE']
-------------------------
Number of Categorical features:  16
```

Our 8 Models:

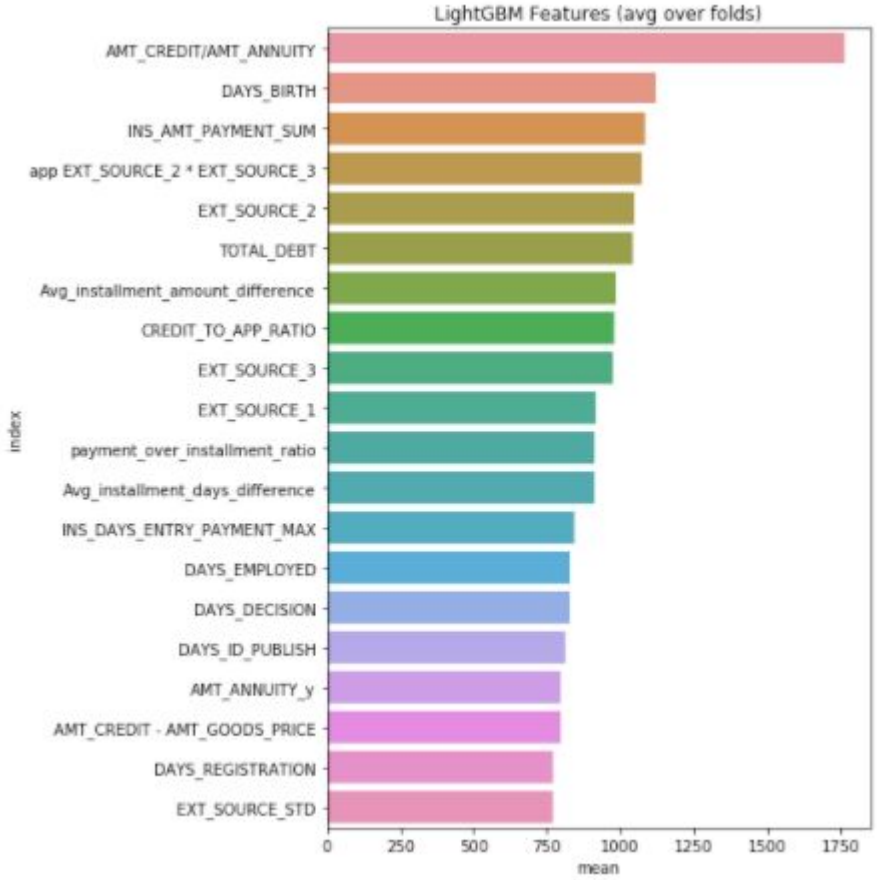
- Baseline Logistic Regression
- Logistic Regression
- Logistic Regression with GridSearchCV
- Random Forest
- XGBoost
- LightGBM
- LightGBM with KFolds
- Deep Learning

Models Results

Evaluation Metrics: AUC-ROC

	exp_name	Train Acc	Valid Acc	Test Acc	Train AUC	Valid AUC	Test AUC	Train Time
0	Baseline_14_features	0.9198	0.9192	0.9159	0.7372	0.7381	0.7383	2.2682
1	LogisticRegression_116_features	0.9200	0.9200	0.9162	0.7646	0.7621	0.7634	9.5147
2	LogisticRegression_GSCV_116_features	0.9200	0.9200	0.9162	0.7646	0.7621	0.7634	346.6979
3	RamdomForest_116_features	0.9198	0.9194	0.9160	0.7507	0.7484	0.7491	56.4632
4	XGBoost_116_features	0.9287	0.9193	0.9152	0.8833	0.7733	0.7774	50.3208
5	LGBM_116_features	0.9430	0.9193	0.9150	0.9716	0.7715	0.7757	19.9208
6	LGBM_Advanced_116_features	0.9589	0.9191	0.9160	0.9907	0.7678	0.7779	19.9208
7	Deep_Learning_116_features	0.9201	0.9197	0.9160	0.8207	0.7677	0.7712	19.9208

Feature Importance for LightGBM (Top 20)



Model Results on Kaggle

Submission and Description	Private Score	Public Score	Use for Final Score
LGBM_submission.csv 4 minutes ago by Neelan Scheumann add submission details	0.77474	0.78264	☐